APPLICATION

FOR

UNITED STATES LETTERS PATENT

TITLE: **DATA STORAGE SYSTEM**

APPLICANT: **JOHN K. WALTON AND DANIEL CASTEL**

"EXPRESS MAIL" Mailing Label Number EL010081281US

Date of Deposit December 30, 1998

I hereby certify under 37 CFR 1.10 that this correspondence is being deposited with the United States Postal Service as **"Express Mail Post Office To Addressee"** with sufficient postage on the date indicated above and is addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231.

JCM

Jason C. Morrissey

UNITED STATES PATENT APPLICATION

of

John K. Walton, Daniel Castel and Kendell Alan Chilton

for

DATA STORAGE SYSTEM

Daly, Crowley & Mofford
275 Turnpike Street, Suite 101
Canton, Massachusetts 02021-2310
Telephone (781) 401-9988
Facsimile (781) 401-9966

Express Mail Label No. ET806886545US

DATA STORAGE SYSTEM

Background of the Invention

This invention relates generally to data storage systems, and more particularly to data storage systems having redundancy arrangements to protect against total system failure in the event of a failure in a component or subassembly of the storage system.

As is known in the art, large host computer systems require large capacity data storage systems. These large computer systems generally includes data processors which perform many operations on data introduced to the computer system through peripherals including the data storage system. The results of these operations are output to peripherals, including the storage system.

One type of data storage system is a magnetic disk storage system. Here a bank of disk drives and the computer system are coupled together through an interface. The interface includes CPU, or "front end", directors (or controllers) and "back end" disk directors (or controllers). The interface operates the directors in such a way that they are transparent to the computer. That is, data is stored in, and retrieved from, the bank of disk drives in such a way that the computer system merely thinks it is operating with one large memory. One such system is described in U.S. Patent 5,206,939, entitled "System and Method for Disk Mapping and Data Retrieval", inventors Moshe Yanai, Natan Vishlitzky, Bruno Alterescu and Daniel Castel, issued April 27, 1993, and assigned to the same assignee as the present invention.

As described in such U.S. Patent, the interface may also include, in addition to the CPU directors and disk directors, addressable cache memories. The cache memory is a semiconductor memory and is provided to rapidly store data

from the computer system before storage in the disk drives, and, on the other hand, store data from the disk drives prior to being sent to the computer. The cache memory being a semiconductor memory, as distinguished from a magnetic memory as in the case of the disk drives, is much faster than the disk drives in reading and writing data.

The CPU directors, disk directors and cache memory are interconnected through a backplane printed circuit board. More particularly, disk directors are mounted on disk director printed circuit boards. CPU directors are mounted on CPU director printed circuit boards. And, cache memories are mounted on cache memory printed circuit boards. The disk director, CPU director and cache memory printed circuit boards plug into the backplane printed circuit board. In order to provide data integrity in case of a failure in a director, the backplane printed circuit board has a pair of buses. One set of the disk directors is connected to one bus and another set of the disk directors is connected to the other bus. Likewise, one set the CPU directors is connected to one bus and another set of the CPU directors is connected to the other bus. The cache memories are connected to both buses. Each one of the buses provides data, address and control information. Thus, the use of two buses provides a degree of redundancy to protect against a total system failure in the event that the directors, or disk drives connected to one bus fail and also increases the bandwidth of the system compared with a system which uses a single bus. One such dual bus system is shown in FIG. 1.

In operation, when the host computer wishes to store end-user (i.e., host computer) data at an address, the host computer issues a write request to one of the front-end directors to perform a write command. One of the front-end directors replies to the request and asks the host computer

for the data. After the request has passed to the requesting one of the front-end directors, the director determines the size of the end-user data and reserves space in the cache memory to store the request. The front-end director then produces control signals on either one of the busses connected to such front-end director. The host computer then transfers the data to the front-end director. The front-end director then advises the host computer that the transfer is complete. The front-end director looks up in a Table, not shown, stored in the cache memory to determine which one of the rear-end directors is to handle this request. The Table maps the host computer address into an address in the bank of disk drives. The front-end director then puts a notification in a "mail box" (not shown and stored in the cache memory) for the rear-end director which is to handle the request, the amount of the data and the disk address for the data. Other rear-end directors poll the cache memory when they are idle to check their "mail boxes". If the polled "mail box" indicates a transfer is to be made, the rear-end director processes the request, addresses the disk drive in the bank, reads the data from the cache memory and writes it into the addresses of a disk drive in the bank. When end-user data previously stored in the bank of disk drives is to be read from the disk drive and returned to the host computer, the interface system operates in a reciprocal manner. The internal operation of the interface, (e.g. "mail-box polling", event flags, data structures, device tables, queues, etc.) is controlled by interface state data which passes between the directors through the cache memory. Further, end-user data is transferred through the interface as a series of multi-word transfers, or bursts. Each word transfer in a multi-word transfer is here, for example, 64 bits. Here, an end-user

data transfer is made up of, for example, 32 bursts. Each interface state word is a single word having, for example, 64 bits.

It is first noted that the end-user data and interface state data are transferred among the directors and the cache memory on the busses. The transfer of each word, whether a burst of end-user data or an interface state data passes through the interface in the same manner; i.e., requiring a fixed amount of overhead, i.e., bus arbitration, etc. Each one of the two busses must share its bandwidth with both end-user data and the interface state data. Therefore, the bandwidth of the system may not be totally allocated to end-user data transfer between the host computer and the bank of disk drives.

Summary of the Invention

In accordance with the present invention, a data storage system is provided wherein end-user data is transferred between a host computer and a bank of disk drives through an interface. The interface includes a memory and a plurality of directors interconnected through an interface state data bus and a plurality of end-user data busses. At least one front-end one of the directors is in communication with the host computer and at least one rear-end one of the directors is in communication with the bank of disk drives. The interface state data bus section is in communication with: both the at least one front-end one and the at least one rear-end one of the directors; and to the memory. Each one of the plurality of end-user data buses has a first end coupled to a corresponding one of the plurality of directors and a second end coupled to the memory. The plurality of directors control the end-user data transfer between the host computer and the bank of disk drives through the memory in response to interface state

data generated by the directors as such end-user data passes through the end-user data busses. The generated interface state data is transferred among the directors through the memory as such end-user data passes through the end-user bus.

With such an arrangement, the system bandwidth is increased because end-user data and interface state data are carried on separate bus systems within the interface.

Brief Description of the Drawing

For a more complete understanding of the invention, reference is now made to the following description taken together in conjunction with the accompanying drawing, in which:

FIG. 1 is a block diagram of a memory system according to the PRIOR ART;

FIG. 2 is a block diagram of a memory system according to the invention;

FIG. 3 is a block diagram of an exemplary one of a plurality of cache memory printed circuit boards used in the system of FIG. 2;

FIG. 4 is a block diagram of an exemplary one of a plurality of front-end directors used in the system of FIG. 2;

FIG. 5 is a block diagram of an exemplary one of a plurality of rear-end directors used in the system of FIG. 2;

FIG. 6 is a block diagram of an exemplary one of a plurality of ASIC control logics used in the cache memories of FIG. 3;

FIG. 7 is a block diagram of another embodiment of an interface in accordance with the invention; and

FIG. 8 is a block diagram of still another embodiment of an interface in accordance with the invention.

Detailed Description

Referring now to FIG. 2, a computer system 100 is shown. The computer system 100 includes a host computer section 112 (e.g., a main frame or open systems computer section) having a plurality of processors, not shown, for processing end-user data. Portions of the processed end-user data are stored in, and retrieved data from, a bank 116 of disk drives through an interface 118. The interface includes a cache memory section 120, here made up of two identical cache memory printed circuit boards 120_0, 120_1, an exemplary one thereof, here memory board 120_0 being shown in, and to be discussed in detail in connection with, FIG. 3. Suffice it to say here, however, that the memory board 120_0 includes an array of DRAMs, here arranged in four memory regions, i.e., memory region A, memory region B. memory region C and memory region D, as shown in FIG. 3 and described in detail in co-pending patent application Serial No. 09/052,268, entitled "Memory System" filed March 31, 1998, inventor John K. Walton, the entire subject matter thereof being incorporated herein by reference.

Referring again to FIG. 2, the interface 118 also includes a plurality of, here eight directors 122_0-122_7. Here, four of the directors, i.e., directors 122_0-122_3, are front-end one of the directors and are coupled to the host computer 112. Here, four of the directors, i.e., directors 122_4-122_7, are rear-end one of the directors and are coupled to the bank of disk drives 116. Each one of the front-end directors 122_0-122_3 is identical in construction, an exemplary one thereof, here front-end director 122_0 being shown in FIG. 4. It is noted that here a cross-bar switch 123 is included to couple each one of a plurality, here four example four, processors, not shown in the host computer 112, to either port P_0 of cache memory 120_0 via serial bus

7

$126_{0,0}$ or port P_0 of cache memory 120_1 via serial bus $126_{0,1}$, as indicated. Here, each one of the serial busses is a four wire bus having a differential pair of receive wires and a differential pair of transmit wires. Likewise, each one of the rear-end directors 122_4-122_7 is identical in construction, an exemplary one thereof, here rear-end director 122_4 being shown in FIG. 5. It is noted that here a cross-bar switch 123 equivalent to that shown in FIG. 4, is included to couple each one of a plurality, here four example four, rows disk drives, not shown, in bank 116, to either port P_4 of cache memory 120_0 via serial bus $126_{4,0}$ or port P_4 of cache memory 120_1 via serial bus $126_{4,1}$, as indicated.

Referring again to FIG. 2, the interface 118 also includes an interface state data bus section 124, here made up of four interface state data parallel, here 72 wire, busses, i.e., bus A, bus B, bus C, and bus D, for carrying interface state data through the interface 118. The interface state data bus section 124 is coupled to: the front-end directors 122_0-122_3, the rear-end directors 122_4-122_7; and all to the cache memory printed circuit boards 120_0, 120_1. Thus, each one of the four busses is a multi-drop bus. The interface 118 also includes a plurality of, here 16, serial end-user data busses $126_{0,0}$-$126_{7,1}$ for carrying end-user data, as indicated. Each one of the plurality of end-user data busses $126_{0,0}$-$126_{7,1}$ has a first end coupled to a corresponding one of the plurality of directors 122_0-122_7 and a second end coupled to the memory section 120. More particularly, and considering director 122_0, such director is coupled to memory bard 122_0 through end-user data bus $126_{0,0}$ and to memory board 122_1 through end-user data bus $126_{0,1}$. Director 122_1 is coupled to memory bard 122_0 through end-user data bus $126_{1,0}$ and to memory

8

board 122_1 through end-user data bus $126_{1,1}$. The other directors are coupled in like manner, for example, director 122_7 is coupled to memory board 122_0 through end-user data bus $126_{7,0}$ and to memory board 122_1 through end-user data bus $126_{7,1}$, as shown. The plurality of directors 122_0-122_7 control the end-user data transfer between the host computer 112 and the bank of disk drives 116 through the memory 120 via the end-user data busses $126_{0,0}$-$126_{7,1}$ in response to interface state data generated by the directors 122_0-122_7. The interface state data is generated by the directors 122_0-122_7 and is transferred among the directors 122_0-122_7 through the memory section 120 via the interface state bus section 124.

An exemplary one of the cache memories 120_0, 120_1, here memory 120_0 is shown in detail in FIG. 3. Such memory section 120_0 includes a plurality of, here four random access memory (RAM) regions (i.e. RAM region A, RAM region B, RAM region C and RAM region D, as shown, and a matrix of rows and columns of control logic sections, here Application Specific Integrated circuits (ASICs), i.e, control logic section ASIC A,A ... control logic section ASIC D,D. Each one of the four columns of control logic section ASICs is coupled to a corresponding one of the interface state data busses A, B, C, and D, respectively, as shown. More particularly, a first column of control logic sections (i.e., ASICs A,A; B,A; C,A and D,A) are coupled to the A bus. A second column of control logic sections (i.e., ASICs A,B; B,B; C,B and D,B) are coupled to the B bus. A third column of control logic sections (i.e., ASICs A,C; B,C; C,C and D,C) are coupled to the C bus. A fourth column of control logic sections (i.e., ASICs A,D; B,D; C,D and D,D) are coupled to the D bus.

Each one of the rows of the control logic sections ASIC A,A ... ASIC D,D is coupled to a corresponding one of the four RAM regions, RAM region A ... RAM region D, via a DATA/CHIP SELECT, as indicated. The first row of ASICs A,A; A,B; A,C; and A,D is coupled to the DATA/CHIP SELECT BUS of RAM region A. The second row of ASICs B,A; B,B; B,C; and B,D is coupled to the DATA/CHIP SELECT BUS of RAM region B. The third row of ASICs C,A; C,B; C,C; and C,D is coupled to the DATA/CHIP SELECT BUS of RAM region C. The fourth row of ASICs D,A; D,B; D,C; and D,D is coupled to the DATA/CHIP SELECT BUS of RAM region D. It should be noted that the control logic sections ASIC A,A ... ASIC D,D in each of the four rows thereof are interconnected through an arbitration bus, not shown, in a manner described in detail in co-pending patent application entitled "Bus Arbitration System", Serial No. 08/996,807, filed December 23, 1997, inventors Christopher S. MacLellan and John K. Walton, assigned to the same assignee as the present invention, the entire subject matter thereof being incorporated in this patent application.

Each one of the rows of the control logic sections ASIC A,A ... ASIC D,D is coupled to a corresponding one of the four RAM regions, RAM region A ... RAM region D, via an MEMORY ADDRESS/CONTROL BUS, as indicated. The first row of ASICs A,A; A,B; A,C; and A,D is coupled to the MEMORY ADDRESS/CONTROL BUS of RAM region A. The second row of ASICs B,A; B,B; B,C; and B,D is coupled to the MEMORY ADDRESS/MEMORY BUS of RAM region B. The third row of ASICs C,A; C,B; C,C; and C,D is coupled to the MEMORY ADDRESS/ CONTROL BUS of RAM region C. The fourth row of ASICs D,A; D,B; D,C; and D,D is coupled to the MEMORY ADDRESS/ CONTROL BUS of RAM region D.

The cache memory 120_0 also includes a coupling node 130 adapted to couple any one of the ports P_0-P_7 to any one of four ports P_A, P_B, P_C, and P_D. Thus, the coupling node 130 adapted to couple any one of the ports P_0-P_7 to any one of the rows of DATA/CHIP SELECT BUSSES, i.e., any one of the ports P_0-P_7 to any one of the four memory regions, A, B, C or D, selectively in response to control signals produced by the ASICs A,A through D,D on a COUPLING NODE CONTROL BUS of each one of the ASICs A,A through D,D. It should be noted that the end-user data is selectively coupled through the coupling node 130 in accordance with routing information fed thereto by the ASICs A,A through D,D in a manner to be described.

The coupling node 130 includes a cross-bar switch section 132 having a plurality of, here two, cross-bar switches 132a, 132b. The coupling node 130 also includes a plurality of, here four, data selectors 134a through 134d. Each one of the cross-bar switches 132a, 132b is a 4x4 cross-bar switch controlled by control signals fed thereto by the ASICs A,A through D,D. Thus, each one of the cross-bar switches 132a, 132b has four input/outputs (here coupled to ports P_0-P_3, P_4-P_7, respectively, as indicated and four output/inputs coupled to a corresponding one of a pair of input/outputs of the four data selectors 134a through 134d.

Each one of such control logic sections ASICs A,A-D,D is identical in construction, an exemplary one thereof, here control logic section ASIC A,A being shown in detail in FIG. 6 to include a control logic 150 having control logic and a buffer memory 152 described in the above-referenced co-pending patent application entitled "TIMING PROTOCOL FOR A DATA STORAGE SYSTEM", inventor John K. Walton, Serial No. 08/996,809, filed December 23, 1997, assigned to the same assignee as the present invention, the entire subject matter

thereof being incorporated herein by reference. The ASIC (A,A) controls transfer of data between the buffer memory 152 and the one of the plurality of buses (i.e., A bus, B bus B, C bus and D bus) coupled to the control logic section ASIC A,A, here bus A. The control logic section ASIC A,A is adapted to produce a control/data bus request for the one of the control/data buses coupled thereto (here RAM region A) and is adapted to effect the transfer in response to a control/data bus grant fed to the control logic section (here ASIC A,A) in accordance with a protocol described in the above-referenced co-pending application entitled "TIMING PROTOCOL FOR A DATA STORAGE SYSTEM", Serial No. 08/996,809, inventor John K. Walton, filed December 23, 1997, the entire subject matter thereof being incorporated herein by reference. The control logic section ASIC A,A also includes a bus arbitration section 153 described in detail in connection with the above referenced patent application entitled "Bus Arbitration System", filed December 23, 1997, inventors Christopher S. MacLellan and John K. Walton. Here, however, the arbitration section also arbitrates for memory contention with one addition bus, the one end-user data serial busses coupled to either port P_A, P_B, P_C, or P_D by the cross-bar switch section 132, in addition to busses A, B., C, and, D. The control logic 150 includes a decoder 157 for decoding the eight chip select signals and one read/write signals on the A bus as described in detail in the above referenced co-pending patent application Serial No. 09/052,268. The decoder 157 produces address, control, and clock for the memory region A on the MEMORY ADDRESS/CONTROL BUS, as indicated.

The routing information is fed to the ASICs A,A through D,D via the interface state data busses A, B, C, and D. Thus, for example, if a burst end-user data is to be

transferred from director 122_3 to memory region A of cache memory 120_0, it is first noted that such director 122_3 is connected to interface state data bus A and therefore the coupling node routing information is for such end-user data transfer is placed by director 122_3 on interface state data bus A. Referring now also to FIGS. 3 and 6, the ASIC A,A control logic, in response to the coupling node routing information on bus A, presents control information on the COUPLING NODE CONTROL BUS coupled thereto to couple port P_3 of the cache memory 120_0 (which is connected to director 122_3 via end-user bus $126_{3,0}$) through cross-bar switch 132a and selector 134a to port P_A.

Referring now to FIG. 7, another embodiment of an interface, here interface 118' is shown. Here, the interface 118' includes 16 directors $122'_0$ - $122'_{15}$ and four cache memory printed circuit boards $120'_0$-$120'_3$ interconnected through parallel, multi-drop busses A, B, C and D for carrying interface state data and sixty-four sets of serial, point-to-point busses $126'_{0,0}$-$126'_{0,3}$ through $126'_{15,0}$-$126'_{15,3}$ for carrying end-user data, as shown. It is noted that here there are four cross-bar switches 132 for each printed circuit board $120'_0$-$120'_3$. Further, here each one of the directors $122'_0$-$122'_{15}$ includes four cross-bar switches 132', each one being a 4x4 cross-bar switch.

Referring now to FIG. 8 another embodiment of an interface, here interface 118" is shown. Such interface 118" includes 16 directors $122"_0$ - $122"_{15}$ and four cache memory printed circuit boards $120"_0$-$120"_3$ interconnected through four parallel, multi-drop busses TH, TL, BH, and BL, such busses being arranged as shown and as described in co-pending patent application entitled "Data Storage System", inventors Daniel Castel, et al., assigned to the same assignee as the present invention, and filed on the same day

as this patent application, the entire subject mater thereof being incorporated herein by reference. As described in such co-patent application, the cache memory is arranged to two set; one set having low address and one set having high addresses. Thus, here cache memory printed circuit boards $120"_0$ and $120"_1$ are included in the high address memory set and cache memory printed circuit boards $120"_2$ and $120"_3$ are included in the low address memory set. Here, however, the parallel, multi-drop busses TH (i.e., top high), TL (i.e., top low), BH (bottom high), and BL (bottom low B) are used to carry interface state words and end-user data is carried by serial, point-to-point busses $126"_{0,0}$ through $126"_{15,3}$, as indicated in FIG. 8.

Other embodiments are within the spirit and scope of the appended claims.

What is claimed is: